MCB128
Final Project
due May 6, 2026

Instructions:

- Project selection should be done by Monday 4/20, 2026
- Discuss project topics with the instructor and TFs as needed. In the end, you should send an email to the instructor (ccing the TFs) describing the project in order to obtain formal approval.
- This project counts as one homework (out of six total) towards 70% of your final grade.
- All tools are available to you...
- ...but the writing and reasoning should be in your own voice.
- You can discuss your project with everybody: TF, classmates, friends,...
- ...but projects are individual. No group projects.

---

1. Topics

   The topic should have some degree of originality, that is, not reproducing something identical provided with the course materials.

   The type of projects is quite flexible:

   (a) Theoretical: a follow up from class on a math aspect to expands on.
   (b) Coding: Apply any of the methods described in class to some dataset of interest to you.
   (c) Divulgation: A slide (or video) presentation of some paper or subject of interest, hopefully where you have reached some original understanding that you want to communicate to the world.

2. Development/Scope

   (a) You need to take into account the limited amount of time provided. Don't do something too trivial, but do not embark on the equivalent of a PhD thesis.
   (b) Depending on the nature of the project, it is ok to reach a certain point, and then describe what would have to be done next, given more time.

3. Format

   The format is pretty flexible. As you do it, consider including some combination of the following items, some of which may apply to some types of projects but not to others

   (a) Motivation that made you pick that project
   (b) Approach that you are going to take
   (c) Hypothesi(e)s that you plan to test
   (d) Include at least one null hypothesis!
   (e) Description of experiments performed
   (f) Summary of results
   (g) Interpretation
   (h) Next to do (or not to do)

Project ideas

Suggested by Shivam

- Take some pathogenic missense mutations from clinvar and introduce them to ESM2. Then use a sparse autoencoder (interPLM) to see if the feature level annotations of the SAE match the pathogenic function of the variant in the protein.
- Compare the couplings from ESM2 to those of direct coupling analysis. Which does better on real double mutant DMS datasets?
- Build a linear model on top of Evo embeddings to do variant effect prediction for noncoding variants
- Try building your own model and test it on the ProteinGym dataset. You don't have to be competitive here
- Compare generated proteins from EVE and diffusion models

Suggested by Louis

- Train an autoencoder to perform data dimensionality reduction on an MD simulation (I can provide data or provide link to a dataset). Does the latent space separate distinct conformations?
- Compare autoencoders, PCA, t-SNE, and UMAPS with a toy dataset
- Show that a linear autoencoder is the same as PCA. How do different nonlinear activations change the subspace?
- Compare a 2nd-order/3rd-order HMM to MLP for protein secondary structure prediction
- Fine-tune ESM2 on the protein secondary structure task. Does fine-tuning the weights of esm2 further improve the test set accuracy? Why? Compare embeddings before and after fine-tuning
- Compare sequence-only structure prediction models to MSA-based structure prediction models. How deep does the MSA need to be to outperform the sequence-only model?
- Take embeddings from different layers of a pretrained protein LM and test how well they predict different properties: secondary structure, disorder, solvent accessibility, or subcellular localization. Which features and layers contain the relevant information? Do the later layers always perform better than the earlier layers?
- Compare AF2s iterative structure refinement intuition with AF3 or RosettaFold diffusion-style. What problem does diffusion solve that AF2s architecture does not? Are there domains where AF2 might be better?
- Investigate the information and covariance structure in multiple sequence alignments and connect them to coevolution and contact prediction.
- Investigate the effects of different position encoding methods (absolute vs relative, rotational, sinusoidal)

Suggested by Armand

- **Geometry of protein LM embedding spaces.** Take ESM2 embeddings across layers for a set of proteins and visualize how the embedding space changes (PCA, t-SNE, UMAP). Do later layers cluster proteins by function or structure better than earlier layers? How does the intrinsic dimensionality change across layers?

- **VAE for scRNA-seq cell type structure.** Train a VAE on a public scRNA-seq dataset (e.g., PBMCs from 10x Genomics). Does the latent space separate cell types? Compare the latent structure to a standard autoencoder and PCA. How does the KL weight affect the learned representation?
- **Noncoding variant effect prediction with Evo.** Take noncoding variants from ClinVar (pathogenic vs. benign) and extract embeddings from Evo. Train a simple linear probe on top to predict pathogenicity. How does this compare to a conservation-based baseline like CADD or PhyloP?

Suggested by Elena

- **Protein to mRNA correspondence** Modify the TransformerSeq2Seq() code from b4 so that the mRNA tokens are triplets (which make all the sense for mRNAs, similar to DNABERT). Then use that model to translate proteins into mRNAs, and then investigate if the cross-attention maps allows you to discover the genetic code.
- **AF2 vs RosettaFold** We studied in detail AF2. David Baker introduced RosettaFold for the same purpose. You could do a comparative study of the two methods. Where are similar? different?

  https://www.science.org/doi/10.1126/science.abj8754
- **Interpretability of pLM** Simon & Zou. InterPLM: discovering interpretable features in protein language models via sparse autoencoders. Nature Methods 2025.

  https://www.nature.com/articles/s41592-025-02836-7
- **About data leakage** Szymborski & Emad. A flaw in using pre-trained pLLMs in protein-protein interaction inference models. Nature Machine Intelligence volume 8, pages197-208 (2026)

  https://www.nature.com/articles/s42256-025-01176-7
- **Explaining how mutations affect AlphaFold predictions**

  https://www.biorxiv.org/content/10.64898/2025.12.30.697132v3
- **EVO2 Decoder** We studied the genome decoder EVO. Here is the next generation, EVO2.

  https://www.biorxiv.org/content/10.1101/2025.02.18.638918v1